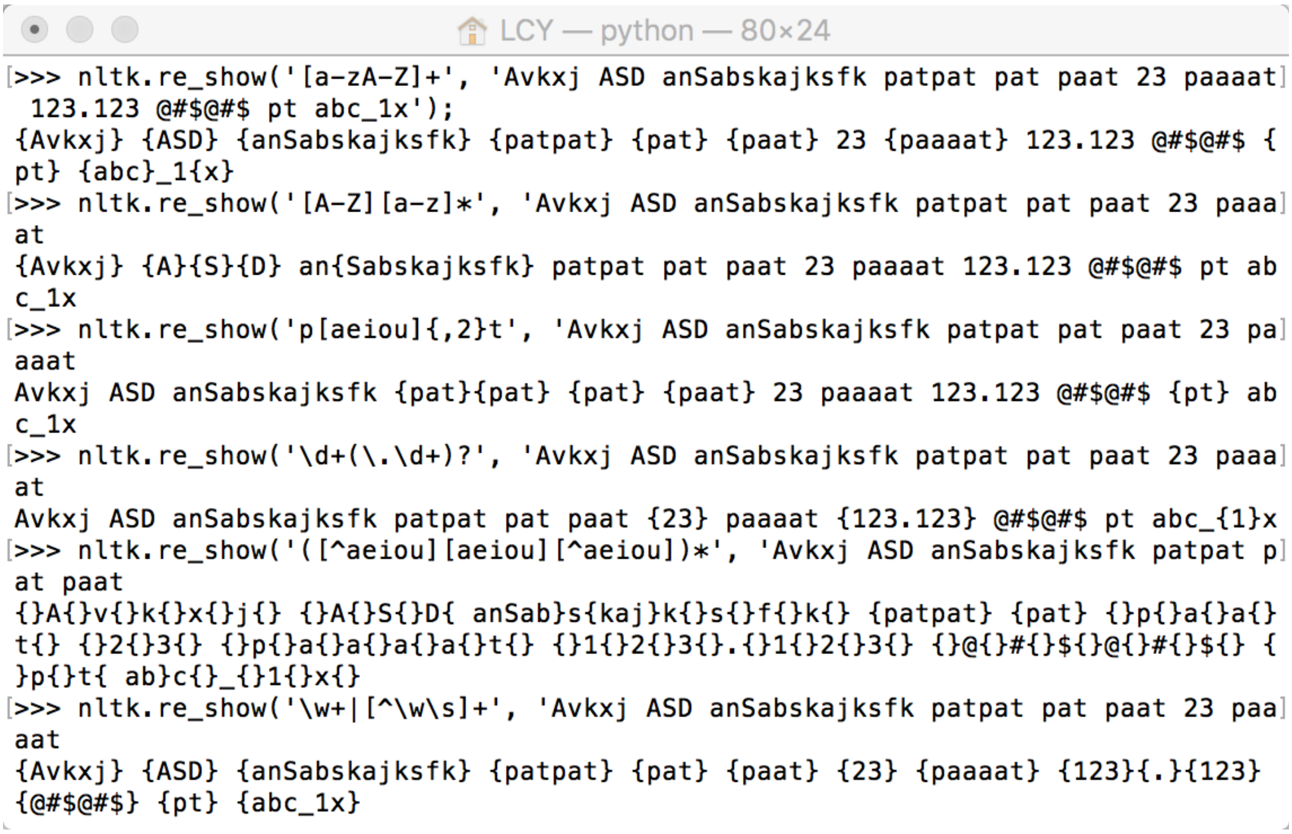
中文信息处理 Chinese Information Processing

第三章 作业

14307130318 刘超颖

1. 说明以下的正则表达式匹配的字符串类： [a-zA-Z]+； [A-Z][a-z]*； p[aeiou]{,2}t； \d+(\.\d+)?； ([^aeiou][aeiou][^aeiou])*； \w+l[^\w\s]+。

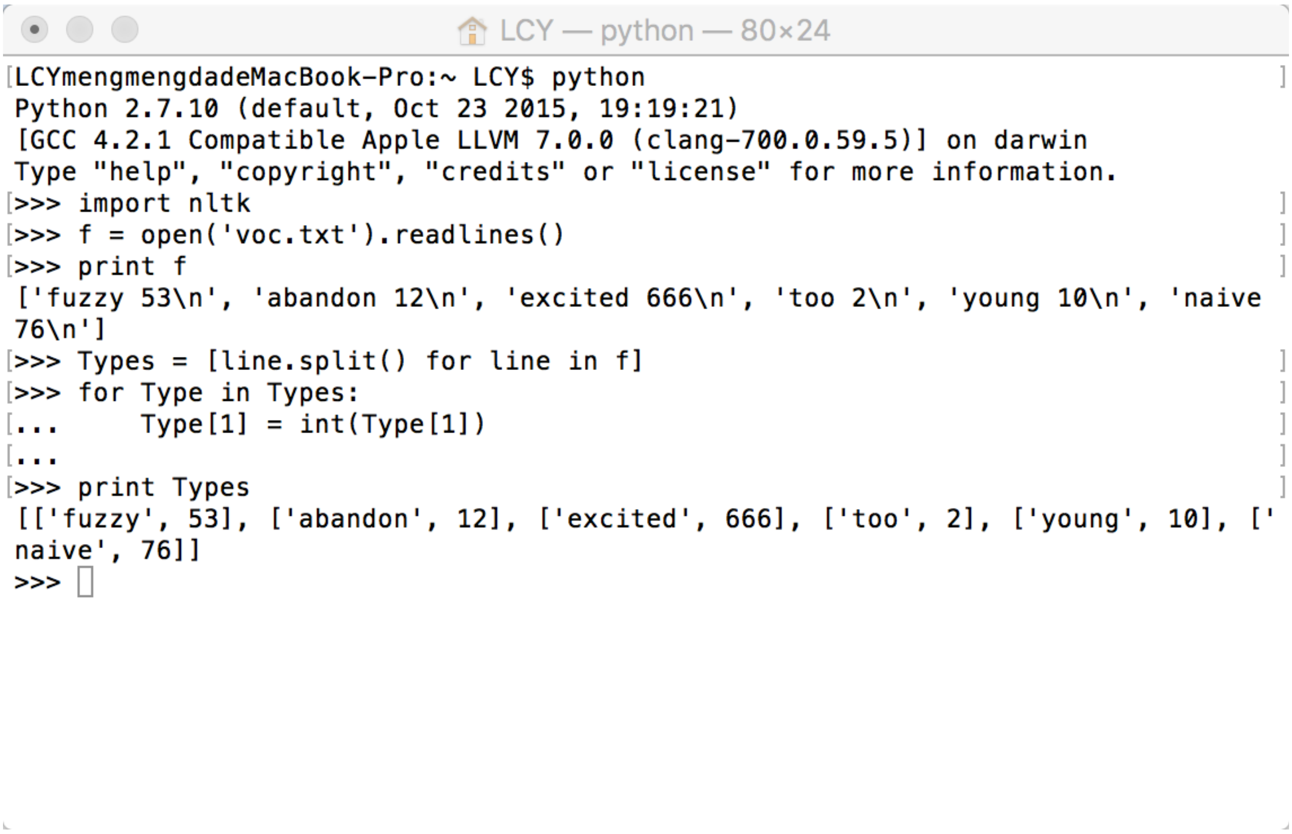
[a-zA-Z]+	匹配	大小写字母组成的字符串，字符串中至少有一个字母
[A-Z][a-z]*		首字母大写其他字母小写的字符串，可以没有小写字母
p[aeiou]{,2}t		首字母为p尾字母为t中间有0-2个元音字母的字符串
\d+(\.\d+)?		一个整数或小数
([^aeiou][aeiou][^aeiou])*		辅原辅形式的三字母单词，重复零次或更多次
\w+l[^\w\s]+		由数字、字母、汉字、下划线组成的字符串或者不存在数字、字母、汉字、下划线、空格的字符串



```
>>> nltk.re_show('[a-zA-Z]+', 'Avkxj ASD anSabskajksfk patpat pat paat 23 paaaat
 123.123 @#$@#$ pt abc_1x');
{Avkxj} {ASD} {anSabskajksfk} {patpat} {pat} {paat} 23 {paaaat} 123.123 @#$@#$ {
pt} {abc}_1{x}
>>> nltk.re_show('[A-Z][a-z]*', 'Avkxj ASD anSabskajksfk patpat pat paat 23 paaa
at
{Avkxj} {A}{S}{D} an{Sabskajksfk} patpat pat paat 23 paaaat 123.123 @#$@#$ pt ab
c_1x
>>> nltk.re_show('p[aeiou]{,2}t', 'Avkxj ASD anSabskajksfk patpat pat paat 23 pa
aaat
Avkxj ASD anSabskajksfk {pat}{pat} {pat} {paat} 23 paaaat 123.123 @#$@#$ {pt} ab
c_1x
>>> nltk.re_show('\d+(\.\d+)?', 'Avkxj ASD anSabskajksfk patpat pat paat 23 paaa
at
Avkxj ASD anSabskajksfk patpat pat paat {23} paaaat {123.123} @#$@#$ pt abc_{1}x
>>> nltk.re_show('([^aeiou][aeiou][^aeiou])*', 'Avkxj ASD anSabskajksfk patpat p
at paat
{}A{}v{}k{}x{}j{} {}A{}S{}D{ anSab}s{kaj}k{}s{}f{}k{} {patpat} {pat} {}p{}a{}a{}
t{} {}2{}3{} {}p{}a{}a{}a{}a{}t{} {}1{}2{}3{}.{}1{}2{}3{} {}@{}#{}${}@{}#{}${} {
}p{}t{ ab}c{}_{}1{}x{}
>>> nltk.re_show('\w+|[^\w\s]+', 'Avkxj ASD anSabskajksfk patpat pat paat 23 paa
aat
{Avkxj} {ASD} {anSabskajksfk} {patpat} {pat} {paat} {23} {paaaat} {123}{.}{123}
{@#$@#$} {pt} {abc_1x}
```

2. 创建一个文件，包含词汇和（任意指定）频率，其中每行包含一个词，一个空格和一个正整数，如：fuzzy 53。使用open(filename).readlines()将文件读入Python 链表。接下来，使用split()将每一行分成两个字段，并使用int()将其中的数字转换为一个整数。结果要求是链表形式：[['fuzzy', 53], ...]。

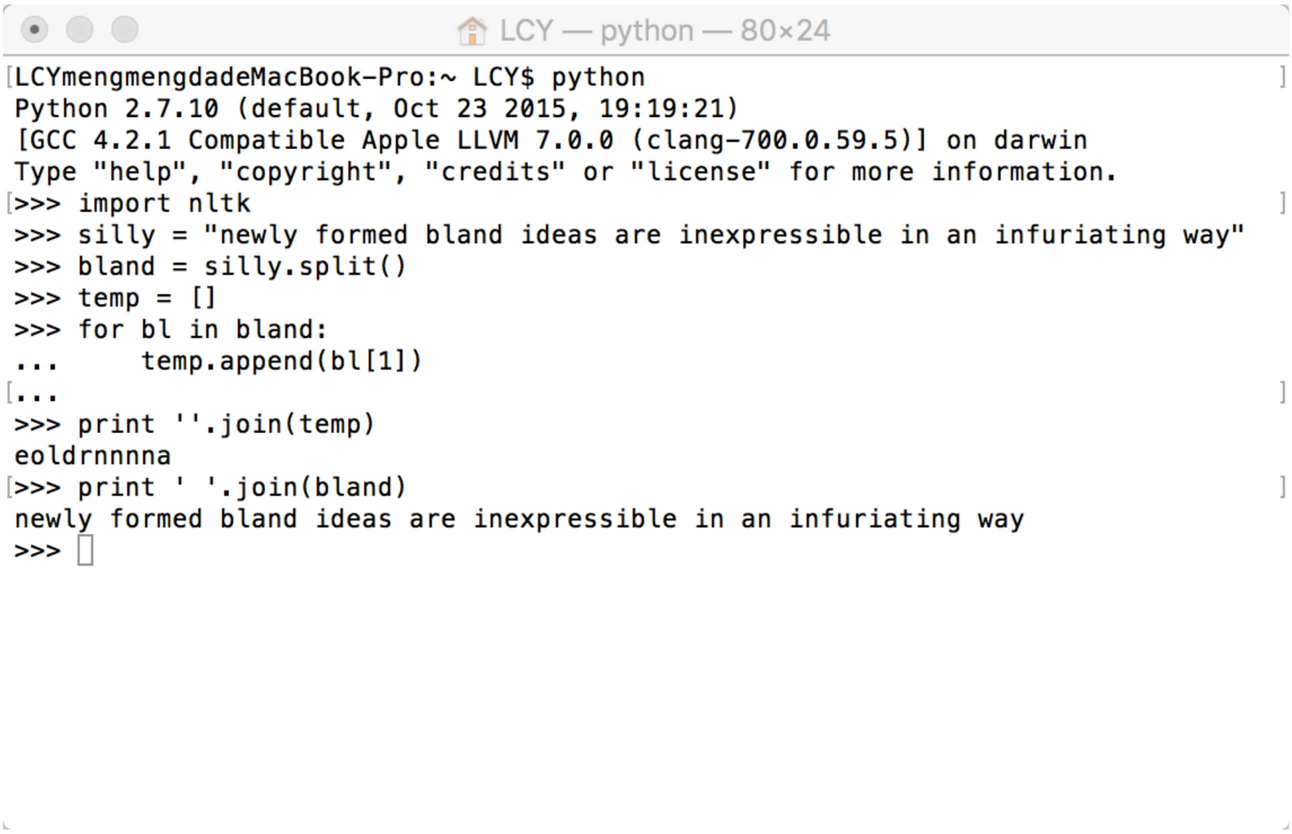
```
>>> f = open('voc.txt').readlines()
>>> Types = [line.split() for line in f]
>>> for Type in Types:
...     Type[1] = int(Type[1])
...
>>> print Types
```



```
LCYmengmengdadeMacBook-Pro:~ LCY$ python
Python 2.7.10 (default, Oct 23 2015, 19:19:21)
[GCC 4.2.1 Compatible Apple LLVM 7.0.0 (clang-700.0.59.5)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> f = open('voc.txt').readlines()
>>> print f
['fuzzy 53\n', 'abandon 12\n', 'excited 666\n', 'too 2\n', 'young 10\n', 'naive
76\n']
>>> Types = [line.split() for line in f]
>>> for Type in Types:
...     Type[1] = int(Type[1])
...
>>> print Types
[['fuzzy', 53], ['abandon', 12], ['excited', 666], ['too', 2], ['young', 10], ['
naive', 76]]
>>> 
```

3. 定义一个变量silly 包含字符串：'newly formed bland ideas are inexpressible in an infuriating way'。编写代码执行以下任务：分割silly 为一个字符串链表，每一个词一个字符串，使用Python 的split()操作，并保存到叫做bland 的变量中；提取silly 中每个词的第二个字母，将它们连接成一个字符串，得到'eoldrnnnna'；使用join()将bland 中的词组合成一个单独的字符串。确保结果字符串中的词以空格隔开。

```
>>> silly = "newly formed bland ideas are inexpressible in an 
infuriating way"
>>> bland = silly.split()
>>> temp = []
>>> for bl in bland:
...     temp.append(bl[1])
...
>>> print ''.join(temp)
>>> print ' '.join(bland)
```



```
LCYmengmengdadeMacBook-Pro:~ LCY$ python
Python 2.7.10 (default, Oct 23 2015, 19:19:21)
[GCC 4.2.1 Compatible Apple LLVM 7.0.0 (clang-700.0.59.5)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> silly = "newly formed bland ideas are inexpressible in an infuriating way"
>>> bland = silly.split()
>>> temp = []
>>> for bl in bland:
...     temp.append(bl[1])
...
>>> print ''.join(temp)
eoldrnnnna
>>> print ' '.join(bland)
newly formed bland ideas are inexpressible in an infuriating way
>>> 
```